Text Analytics

Derrick L. Cogburn
American University
dcogburn@american.edu

Michael J. Hine
Carleton University
mike.hine@carleton.edu

Normand Peladeau
Provalis Research
Peladeau@provalisresearch.com

Victoria Y. Yoon
Virginia Commonwealth U.
vyyoon@vcu.edu

Abstract

This virtual HICSS-54 minitrack recognizes that most global collaboration systems, social media, and information systems of all types, generate enormous amounts of unstructured textual data, including system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this unstructured textual data is readily available, it presents tremendous challenges to researchers trying to analyze these large bodies of text with traditional methods. Text mining in big data analytics is an increasingly important technique for an interdisciplinary group of scholars, practitioners, government officials, and international organizations.

1. Introduction

Building on the success of our annual tutorial on Text Analytics and the corresponding minitrack, we are pleased to introduce the selected papers for the minitrack on Text Analytics at the virtual HICSS-54. Global collaboration systems and information systems of all types generate enormous amounts of unstructured textual data, including system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this unstructured textual data is readily available, it presents tremendous challenges to researchers trying to analyze these large bodies of text with traditional methods. Text mining in big data analytics is an increasingly important technique for an interdisciplinary group of scholars, practitioners, government officials, and international organizations.

2. Minitrack Topics and Themes

The minitrack on Text Analytics is designed to provide an interactive forum for interdisciplinary researchers to discuss the critical issues of text mining and to contribute to the ongoing focus on big data at HICSS. Our minitrack invites papers that apply theoretical and applied text-mining approaches to a wide variety of substantive domains, including, but not limited to:

- Blog posts
- Social media analysis
- Email archives
- Published articles
- Websites
- Meeting transcripts
- Speeches
- Online discussion forums
- Online communities
- Computer logs

And addressing methodological challenges as:

- Automated acquisition and cleaning data
- Working on distributed, high-performance computers
- Overcoming API limitations
- Using LDA, LSA, and other techniques
- Robust Natural Language Processing (NLP) techniques
- Text summarization, classification, and clustering.

As co-chairs of the HICSS Text Analytics minitrack, we are pleased with the growth of our HICSS community. However, the COVID-19 restrictions have hit us very hard this year. We received several great submissions and accepted three excellent papers that highlight various important aspects of this emerging community, including one best paper nomination. Unfortunately, one of our authors withdrew their paper, leaving us with only three to present at HICSS-54. Nonetheless, we are excited about our virtual minitrack, and look forward to discussions of these two papers, and using the remaining time to discuss mechanisms for building this growing and important community. In this next section, we present a brief overview of our two papers, including our Best Paper nomination.

URI: https://hdl.handle.net/10125/70705
978-0-9981331-4-0

3. Paper 1: Automated Topic Analysis for Restricted Scope Health Corpora: Methodology and Comparison with Human Performance

Our first paper is our nomination for Best Paper. It addresses the problem of identifying topics which describe information content, in restricted size sets of scientific papers extracted from publication databases. Conventional computational approaches, based on natural language processing using unsupervised classification algorithms, typically require large numbers of papers to achieve adequate training. The approach presented here uses a simpler word-frequency-based approach coupled with context modeling. An example is provided of its application to corpora resulting from a curated literature search site for COVID-19 research publications. The results are compared with a conventional human-based approach, indicating partial overlap in the topics identified. The findings suggest that computational approaches may provide an alternative to human expert topic analysis, provided adequate contextual models are available.

4. Paper 2: Comparison of Voluntary Versus Mandatory Vaccine Discussions in Online Health Communities: A Text Analytics Approach

Our second paper focuses on the critically important issue of vaccine and sees them as vital health interventions. However, they are controversial, and some people support them while others reject them. Social media discussion and big data are a rich source to understand people's insights about different vaccines and the related topics that concern most of them. This study aims to explore the online discussions about mandatory and voluntary vaccines using text analysis techniques. Reddit social platform is popular in online health discussion and thus data from Reddit is analyzed. The results show that different aspects are discussed for different types of vaccines. The discussion of mandatory vaccines is more interactive and is focused on the risks associated with them. Voluntary vaccines' discussion is focused on their effectiveness and whether to get them or not. The study has important implications for health agencies and researchers as well as for healthcare providers and caregivers.

5. Withdrawn Paper 3: Text Analytics for Business Research

Finally, the paper that was withdrawn addressed text analytics and machine learning techniques as employed increasingly in business research. These techniques have been used to discover or extract new simple features from large and unstructured data. These machine learned features (MLFs) are then used as independent or explanatory variables in the main econometric models for empirical research. Despite this growing trend, there has been little research regarding the impact of using MLFs on statistical inference for empirical research. This paper undertakes parameter estimation issues related to the use of topics/features extracted by Latent Dirichlet Allocation, a popular machine learning technique for text mining. This paper proposes a novel method to extract features that result in the minimum-variance estimation of the regression model parameters. This enables a better use of unstructured text data for econometric modeling in empirical research. The effectiveness of the proposed method is validated with an experimental evaluation study on real-world text data.

6. Towards a Text Mining Community

We believe the Text Mining minitrack has makes an important contribution to HICSS. It has great potential to stimulate the creation of a robust, interdisciplinary text mining research community within HICSS. Given the amount of unstructured textual data generated by widespread collaboration systems and technologies, such a research community would be invaluable. The text mining papers at this 54th Anniversary HICSS represent what we see as an important emergent trend, which we believe will remain for many years to come.